

<https://speakerdeck.com/rmoff/>



Streaming ETL in Practice with PostgreSQL, Apache Kafka, and KSQL

PGConf UK 2018

3 Jul 2018 / Robin Moffatt

@rmoff

robin@confluent.io

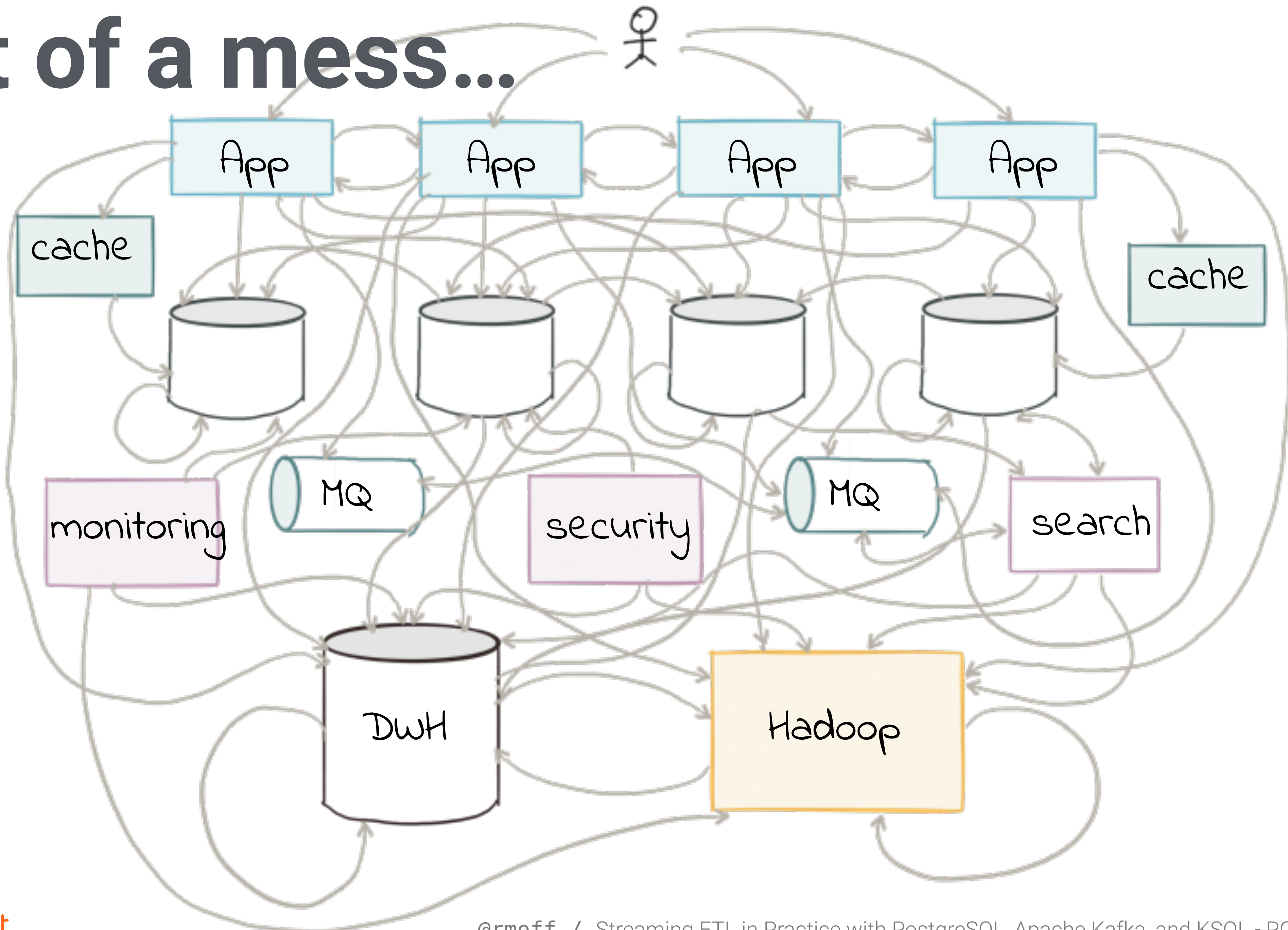
\$ whoami

- Developer Advocate @ Confluent
- Working in data & analytics since 2001
- Oracle ACE Director & Dev Champion 🃏 {🏆}
- Blogging : <http://rmoff.net> & <http://cnfl.io/rmoff>
- Twitter: @rmoff
 - Geek stuff
 - Beer & Fried Breakfasts

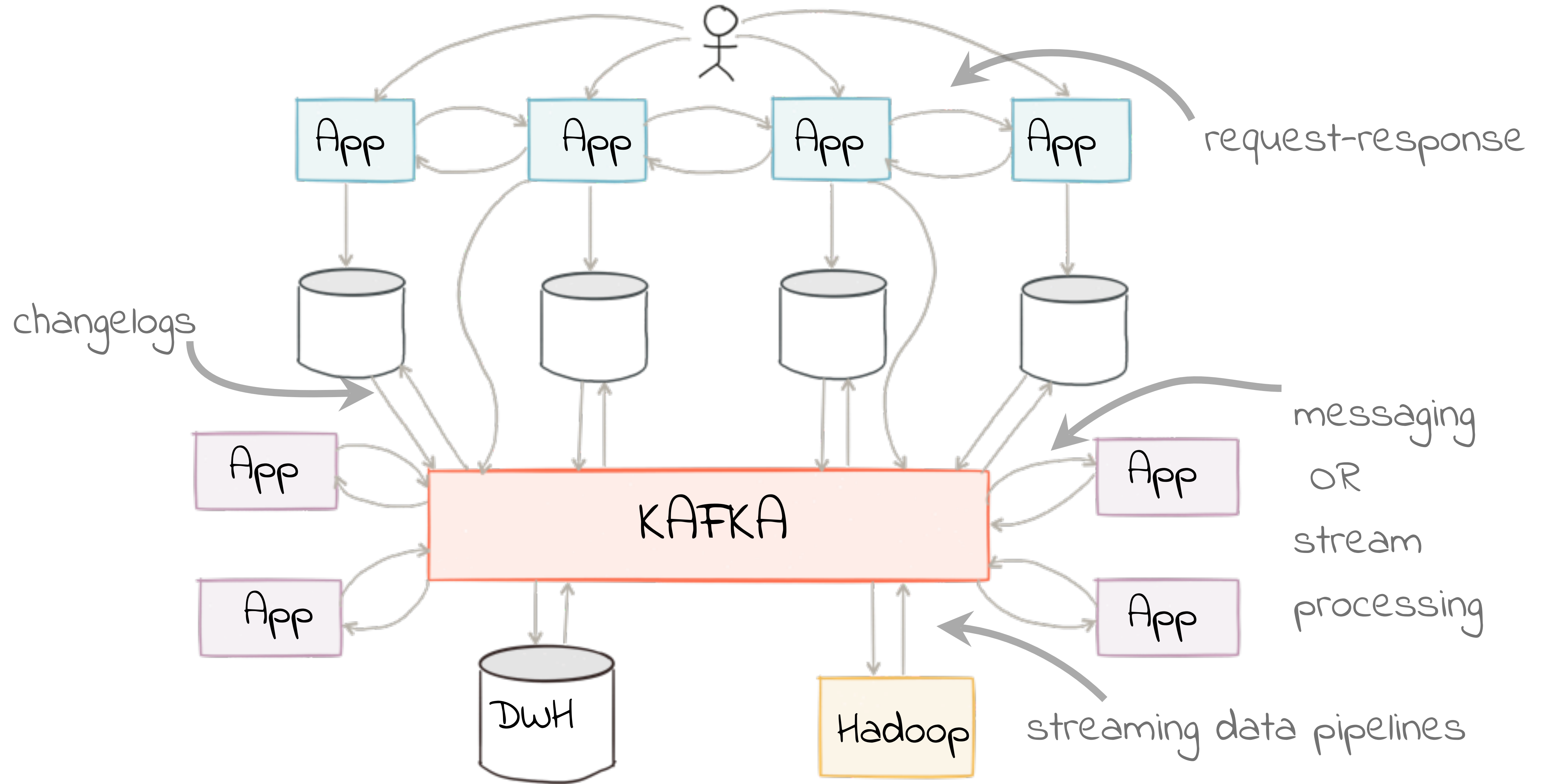
<https://speakerdeck.com/rmoff/>



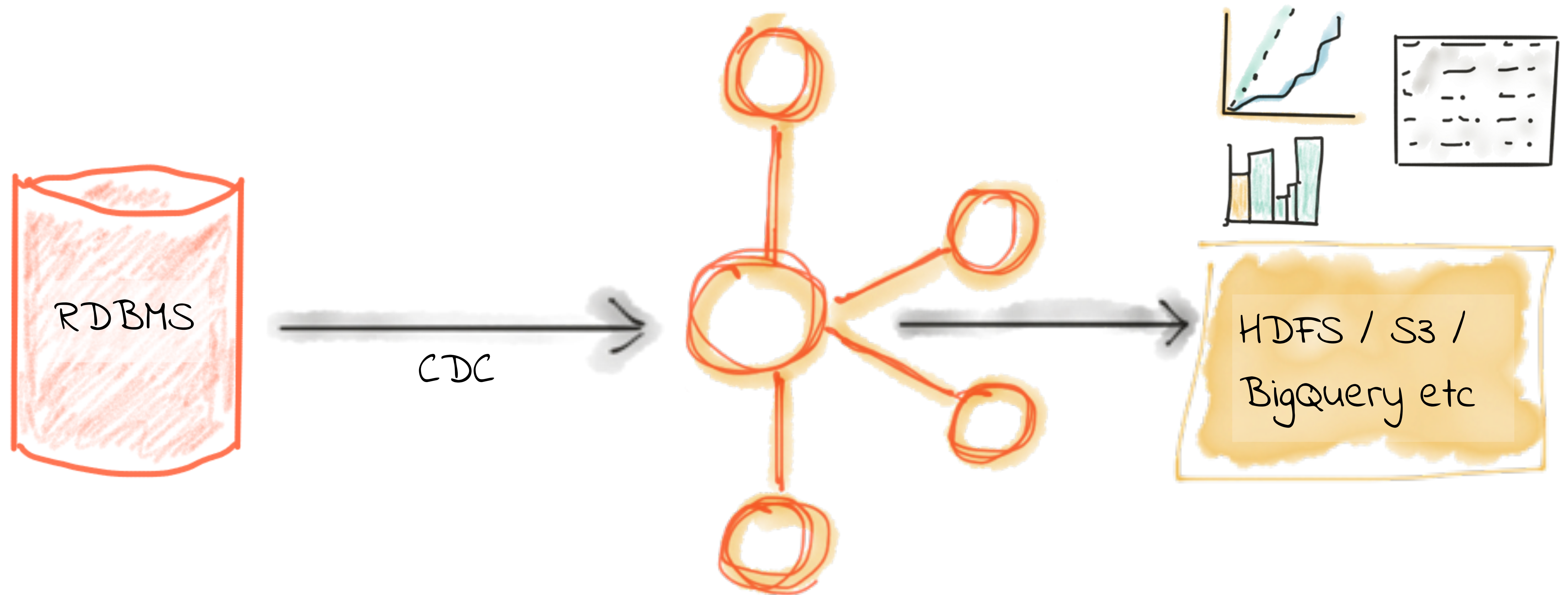
A bit of a mess...



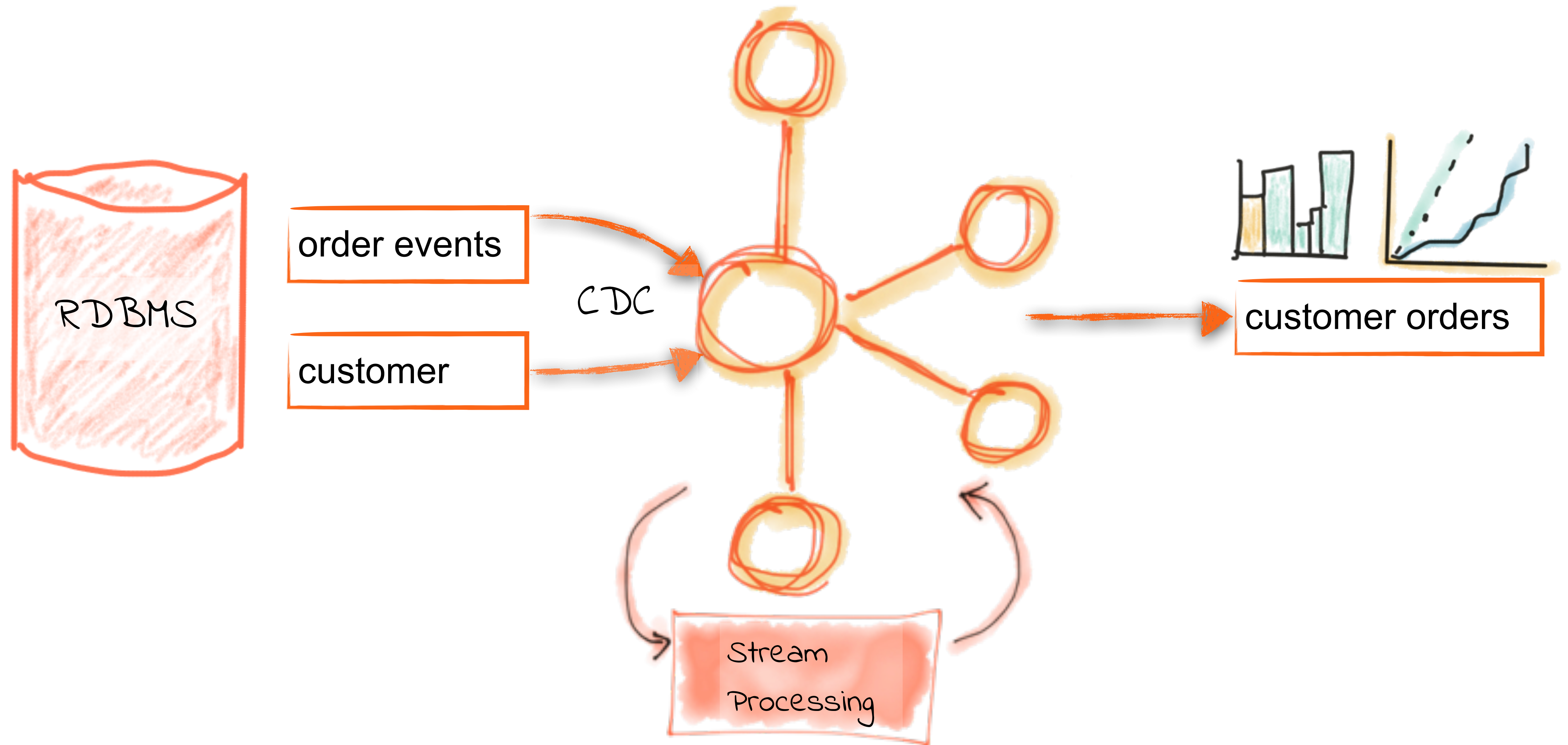
The Streaming Platform



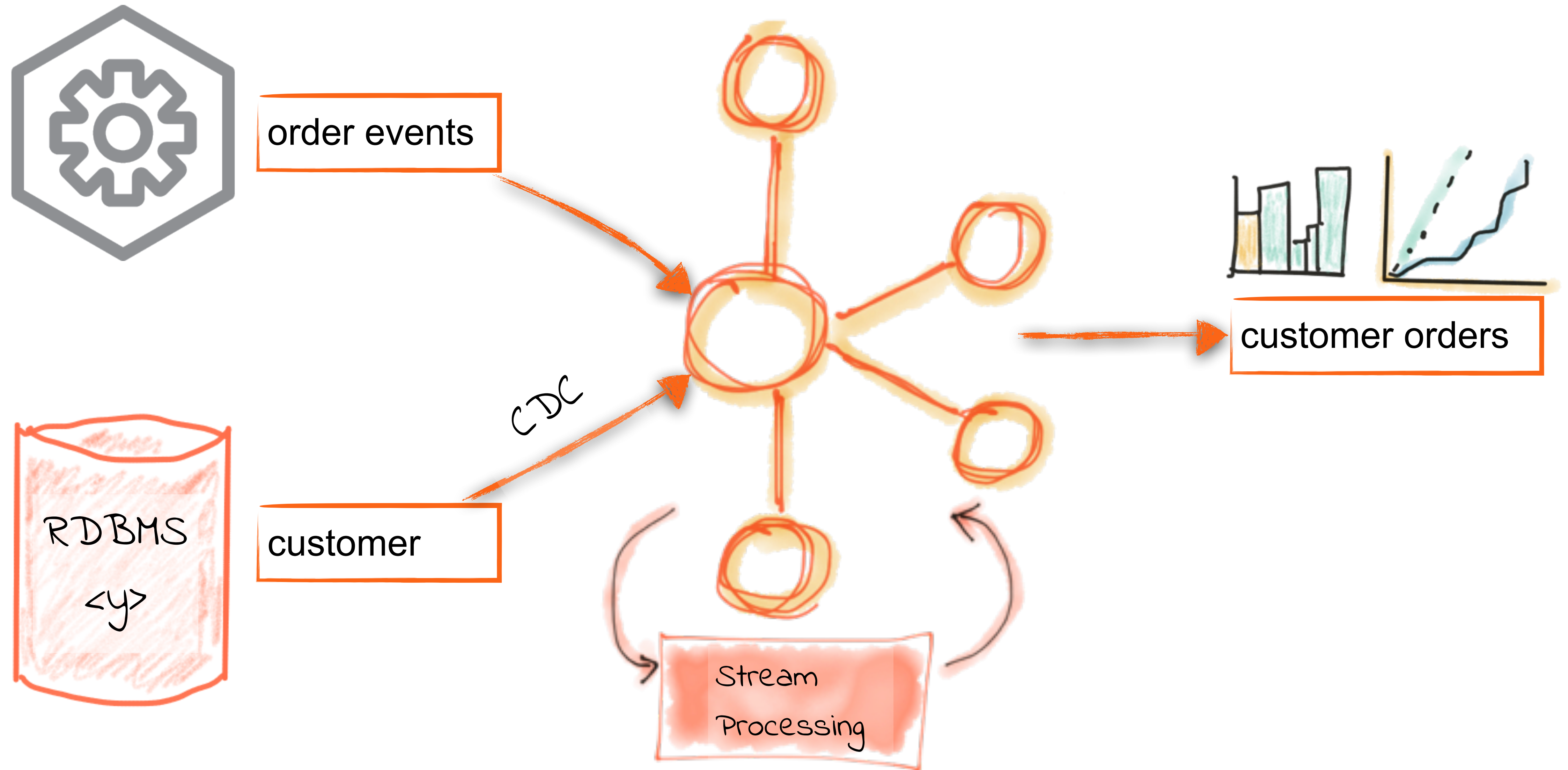
Database offload → Analytics



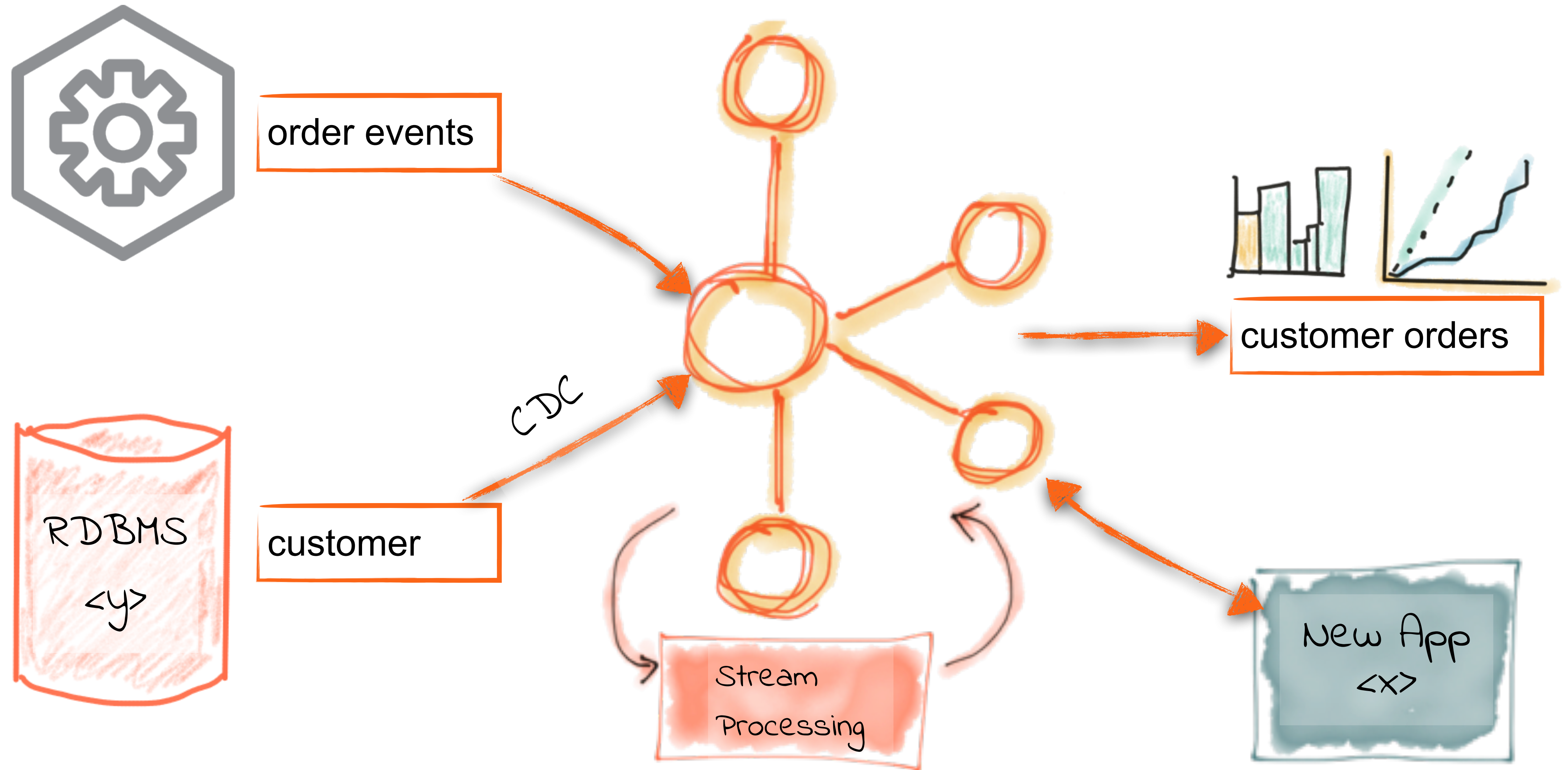
Streaming ETL with Apache Kafka and KSQL



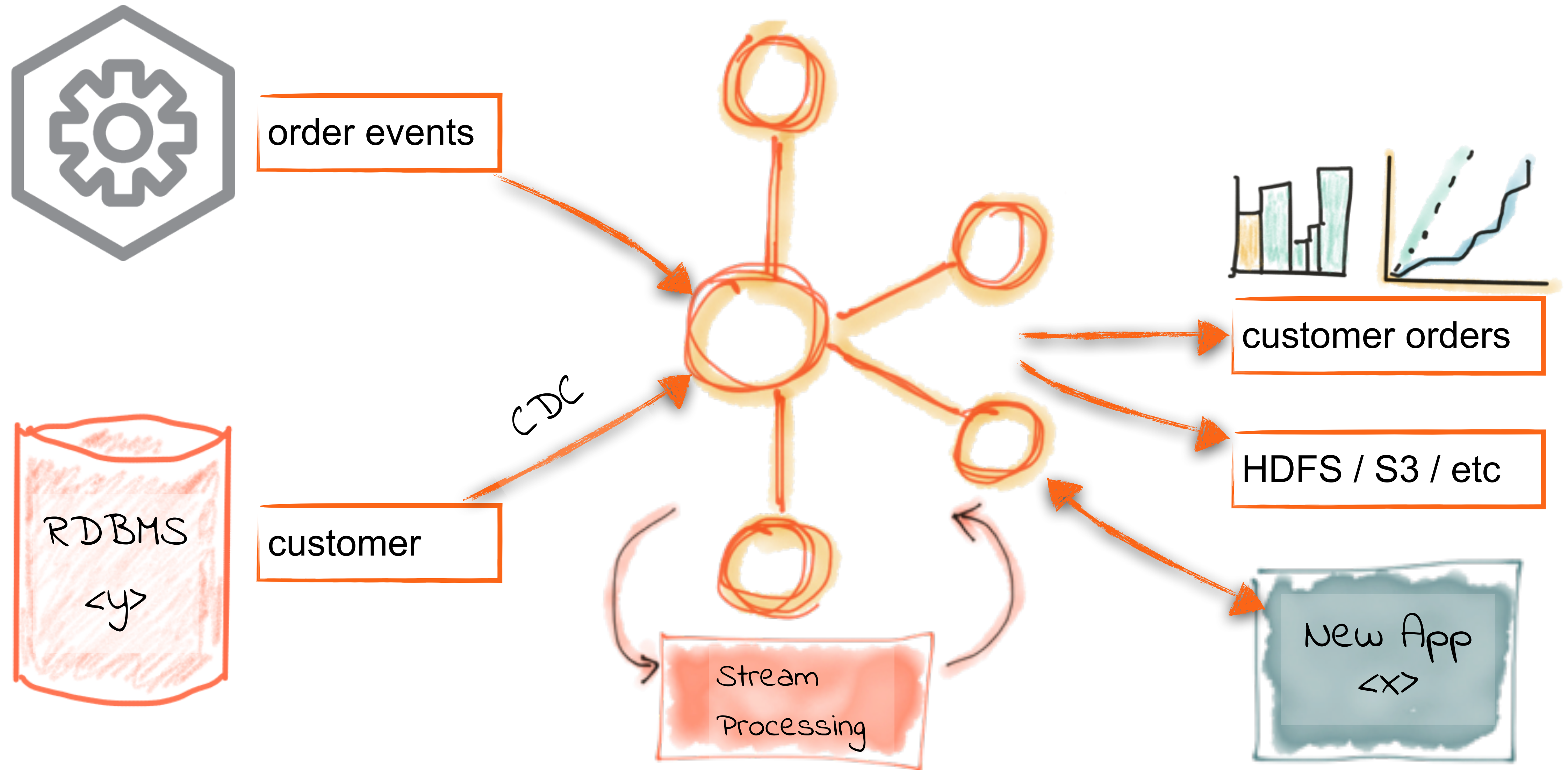
Real-time Event Stream Enrichment



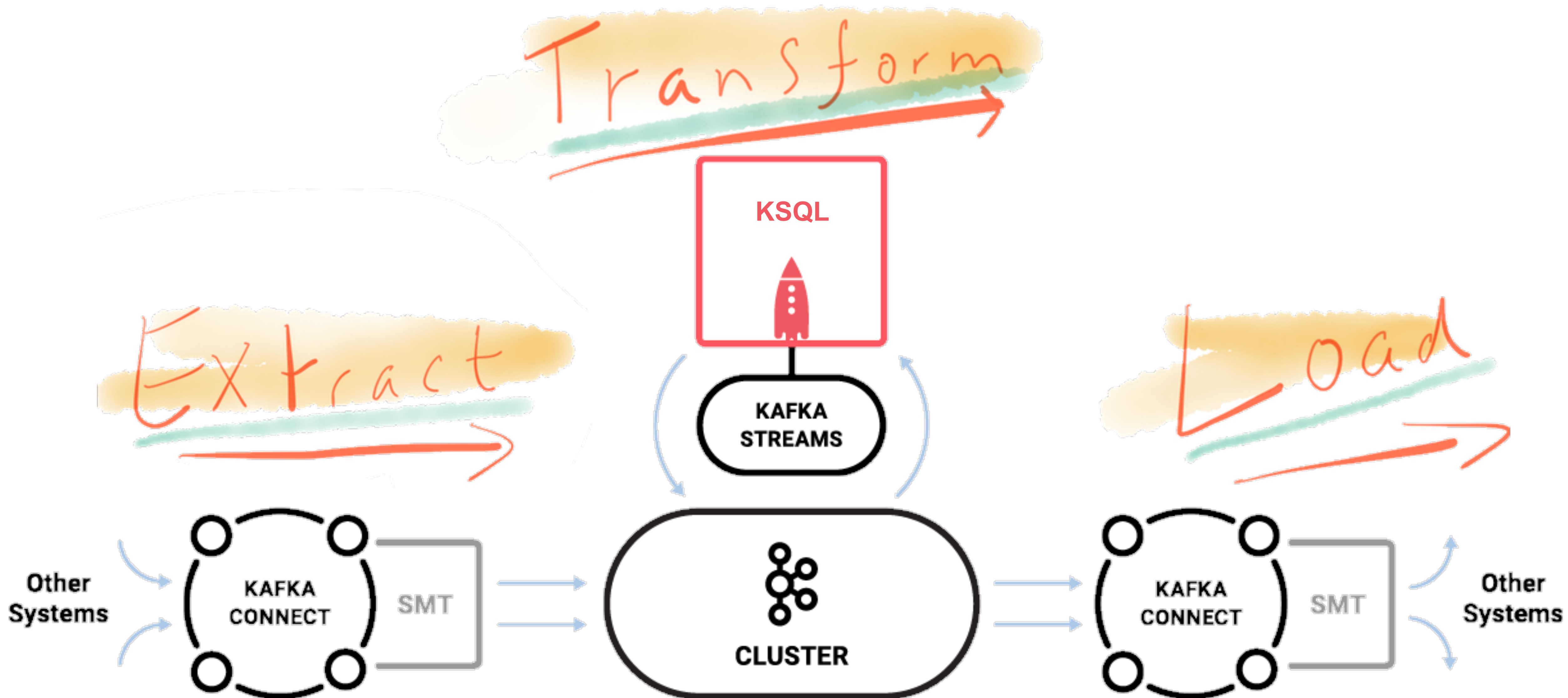
Transform Once, Use Many



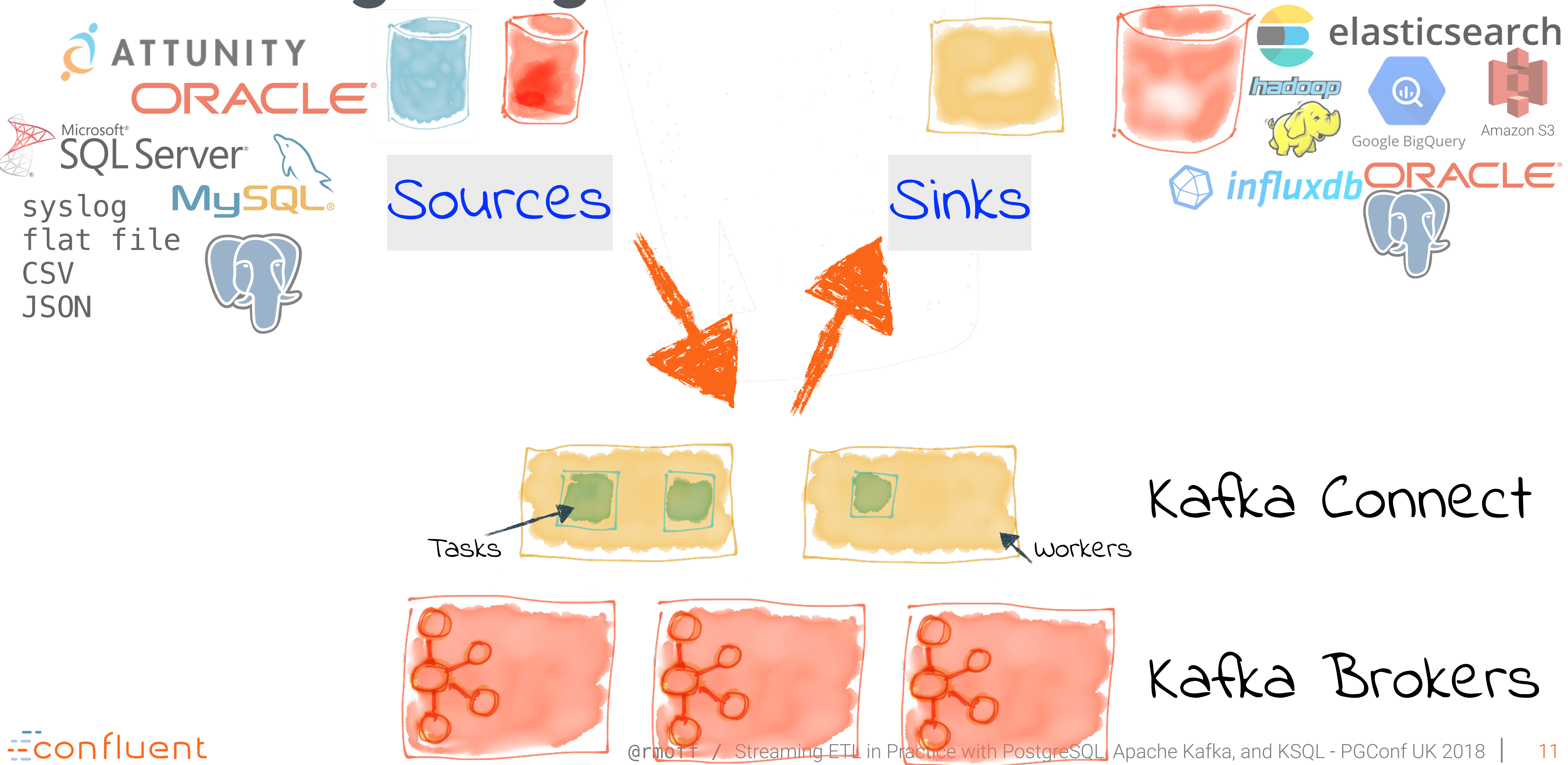
Transform Once, Use Many



Streaming ETL with Apache Kafka



Streaming Integration with Kafka Connect



The Connect API of Apache Kafka®



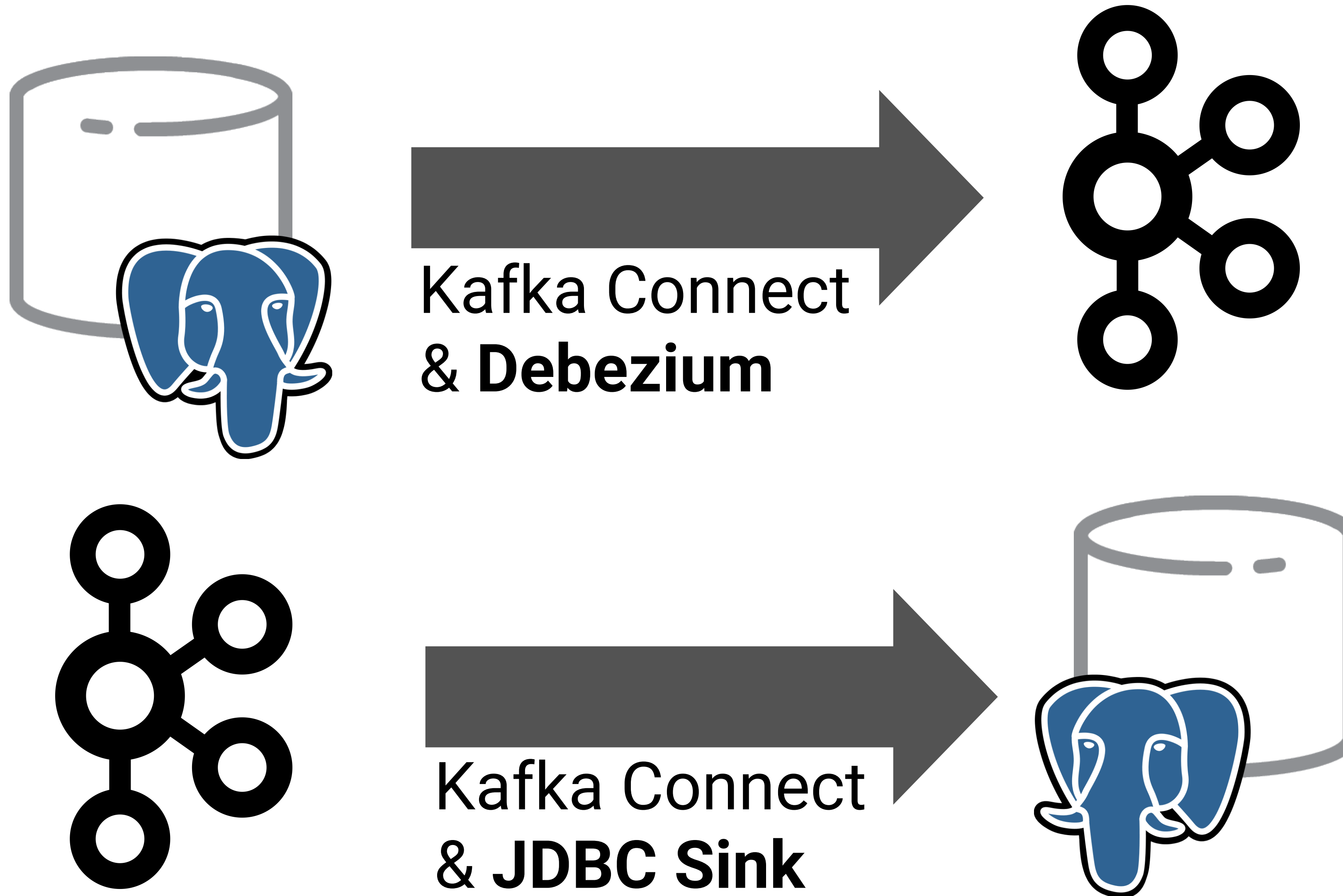
**Reliable and scalable integration of Kafka
with other systems – no coding required.**

- ✓ Centralized management and configuration
- ✓ Support for hundreds of technologies
including RDBMS, Elasticsearch, HDFS, S3
- ✓ Supports CDC ingest of events from RDBMS
- ✓ Preserves data schema
- ✓ Fault tolerant and automatically load balanced
- ✓ Extensible API
- ✓ Single Message Transforms
- ✓ Part of Apache Kafka, included in
Confluent Open Source

```
{  
  "connector.class": "io.confluent.connect.jdbc.JdbcSourceConnector",  
  "connection.url": "jdbc:mysql://localhost:3306/demo?user=rmoff&password=foo",  
  "table.whitelist": "sales,orders,customers"  
}
```

<https://docs.confluent.io/current/connect/>

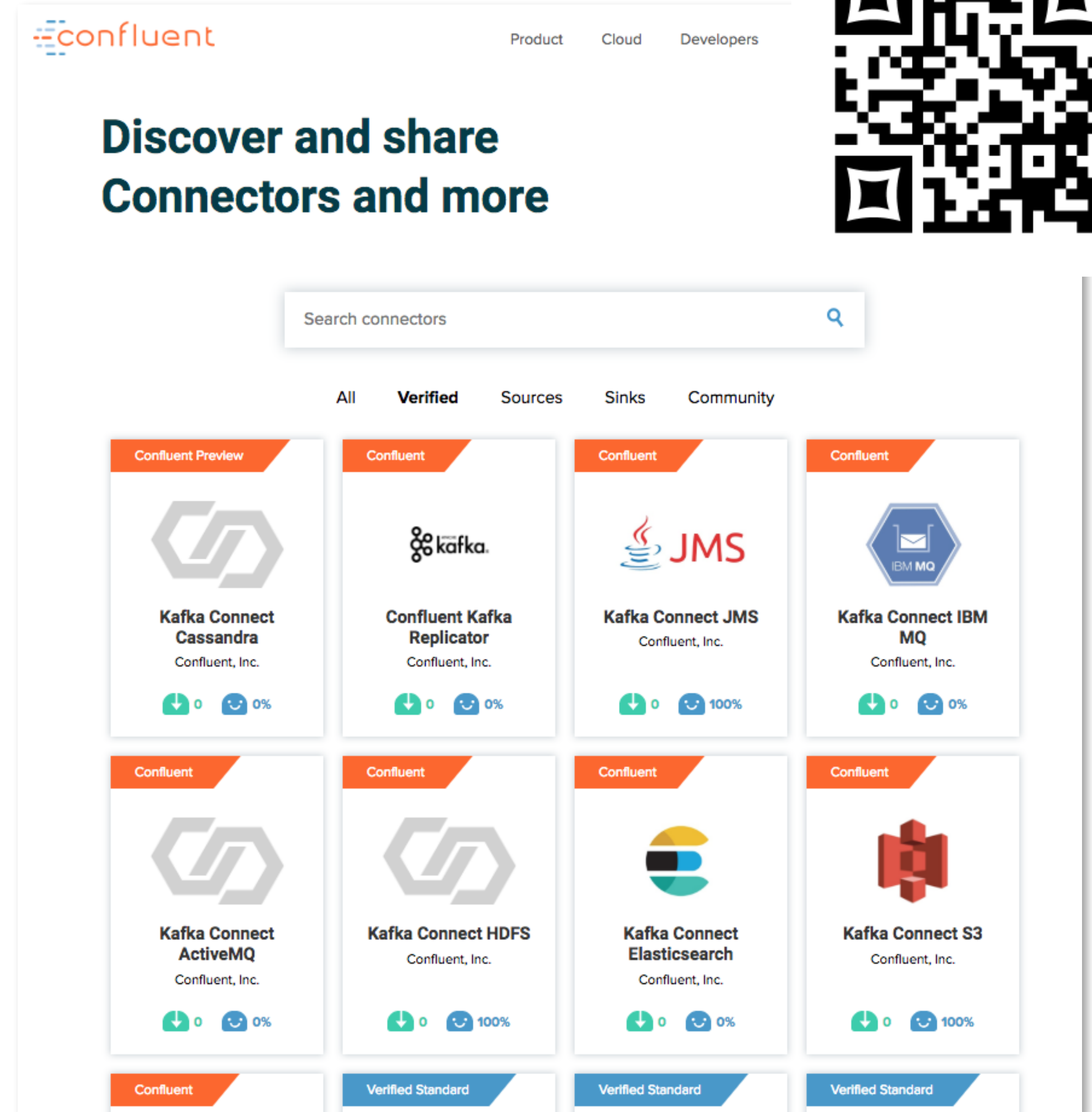
Integrating Postgres with Kafka



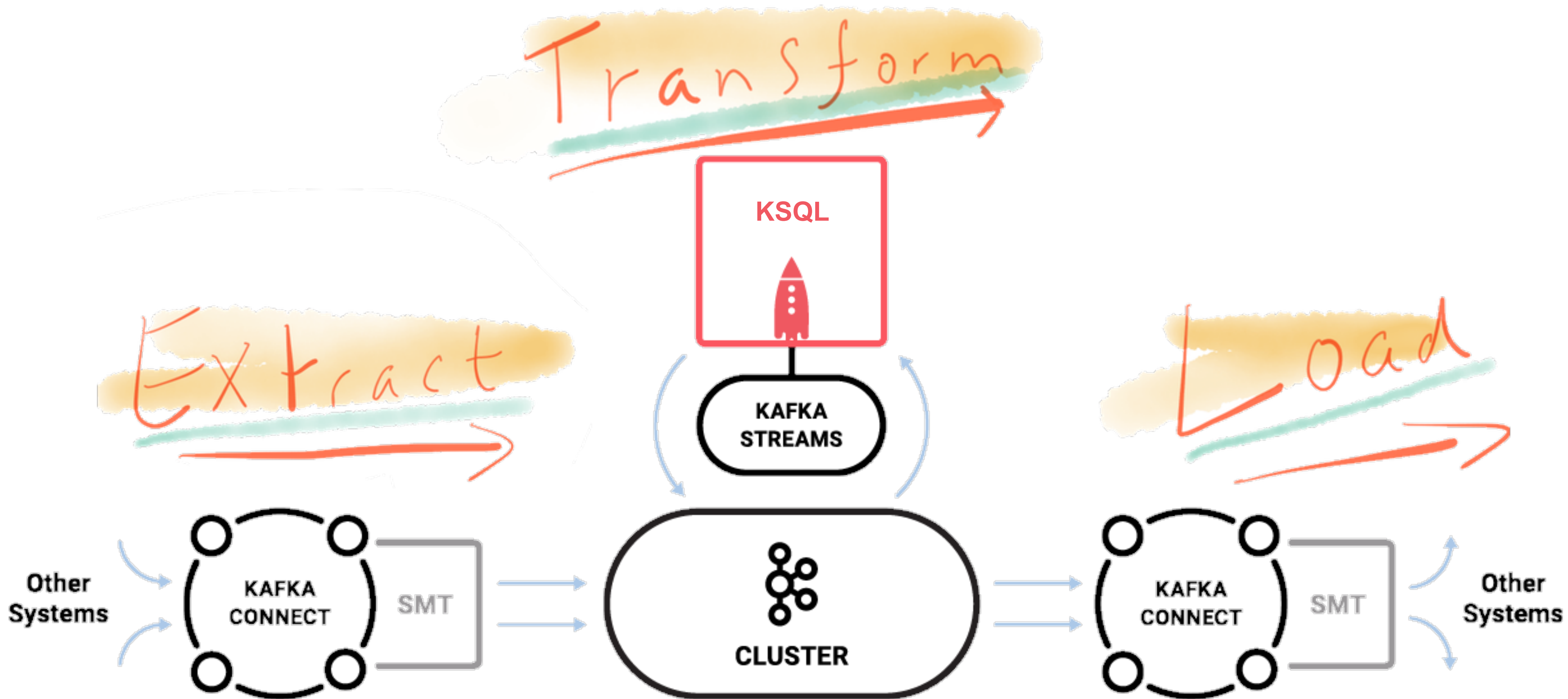
Confluent Hub

- Launched June 2018
- One-stop place to discover and download :
 - Connectors
 - Transformations
 - Converters

hub.confluent.io



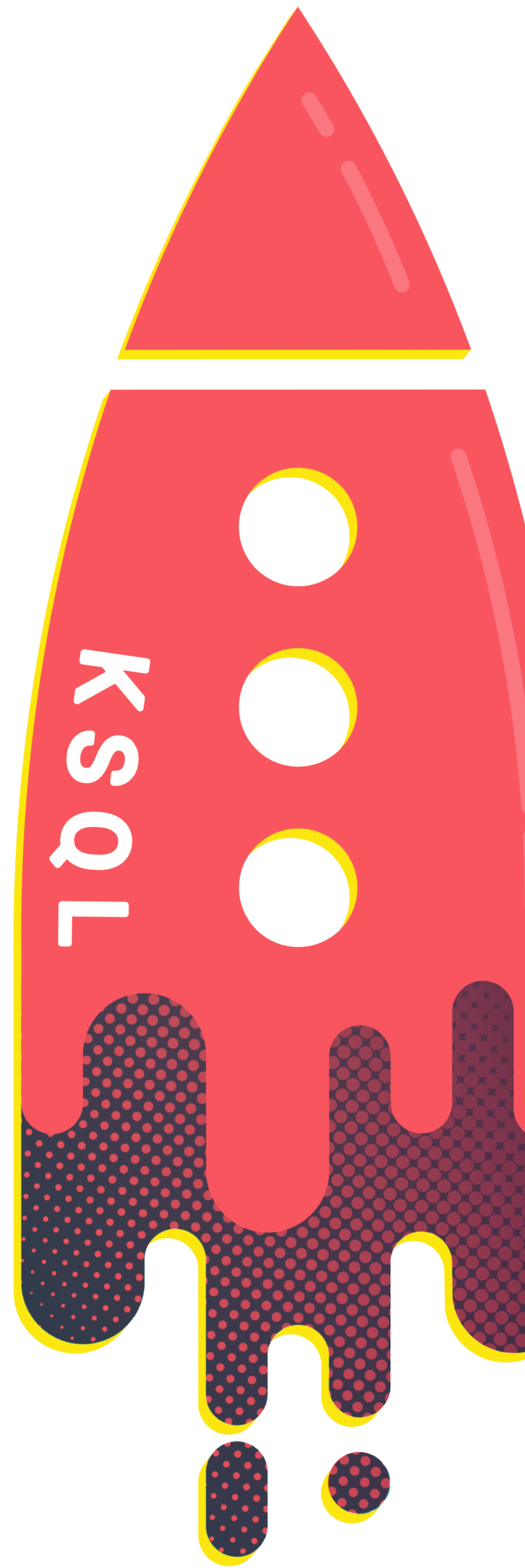
Streaming ETL with Apache Kafka



KSQL

is a

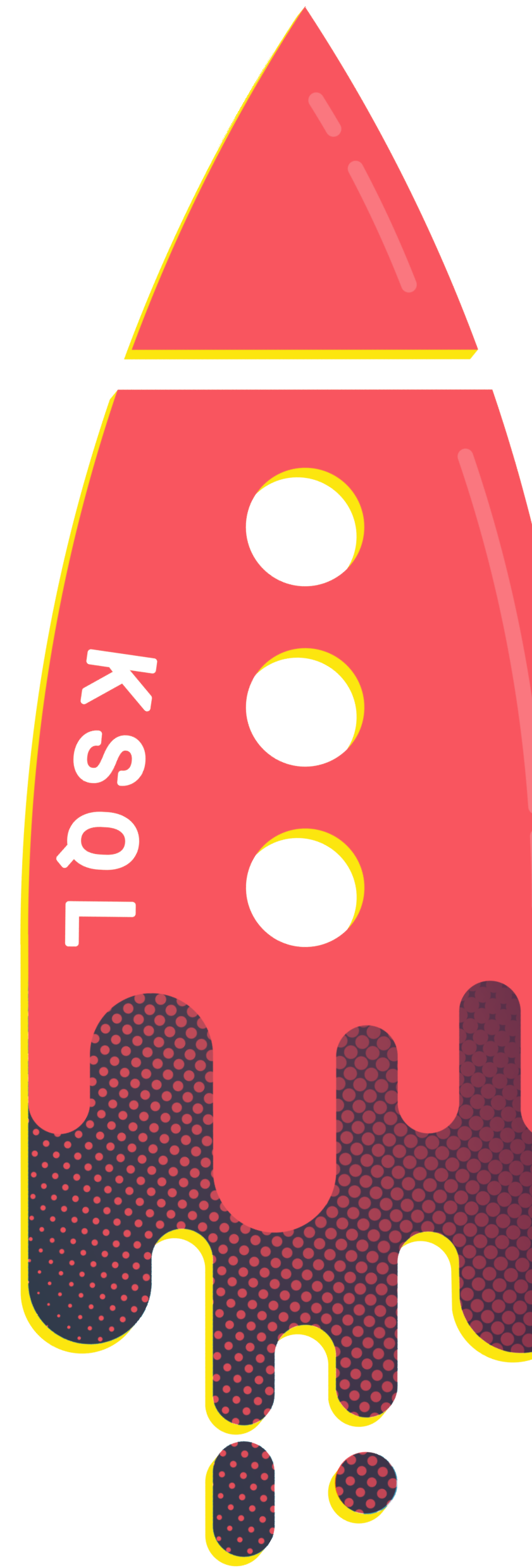
**Declarative
Stream.
Processing
Language**



ksql

is the

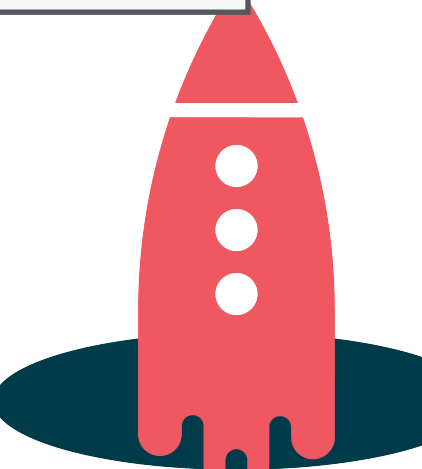
Streaming SQL Engine for Apache Kafka



KSQL for Streaming ETL

Joining, filtering, and aggregating streams of event data

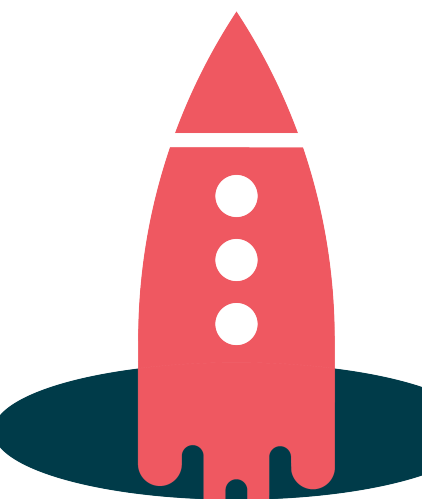
```
CREATE STREAM vip_actions AS
  SELECT userid, page, action
  FROM clickstream c
  LEFT JOIN users u
    ON c.userid = u.user_id
  WHERE u.level = 'Platinum';
```



KSQL for Anomaly Detection

Identifying patterns or anomalies in real-time data,
surfaced in milliseconds

```
CREATE TABLE possible_fraud AS
  SELECT card_number, count(*)
    FROM authorization_attempts
  WINDOW TUMBLING (SIZE 5 SECONDS)
  GROUP BY card_number
  HAVING count(*) > 3;
```

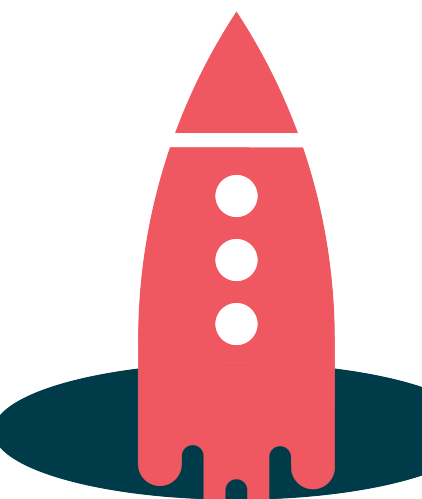


KSQL for Real-Time Monitoring

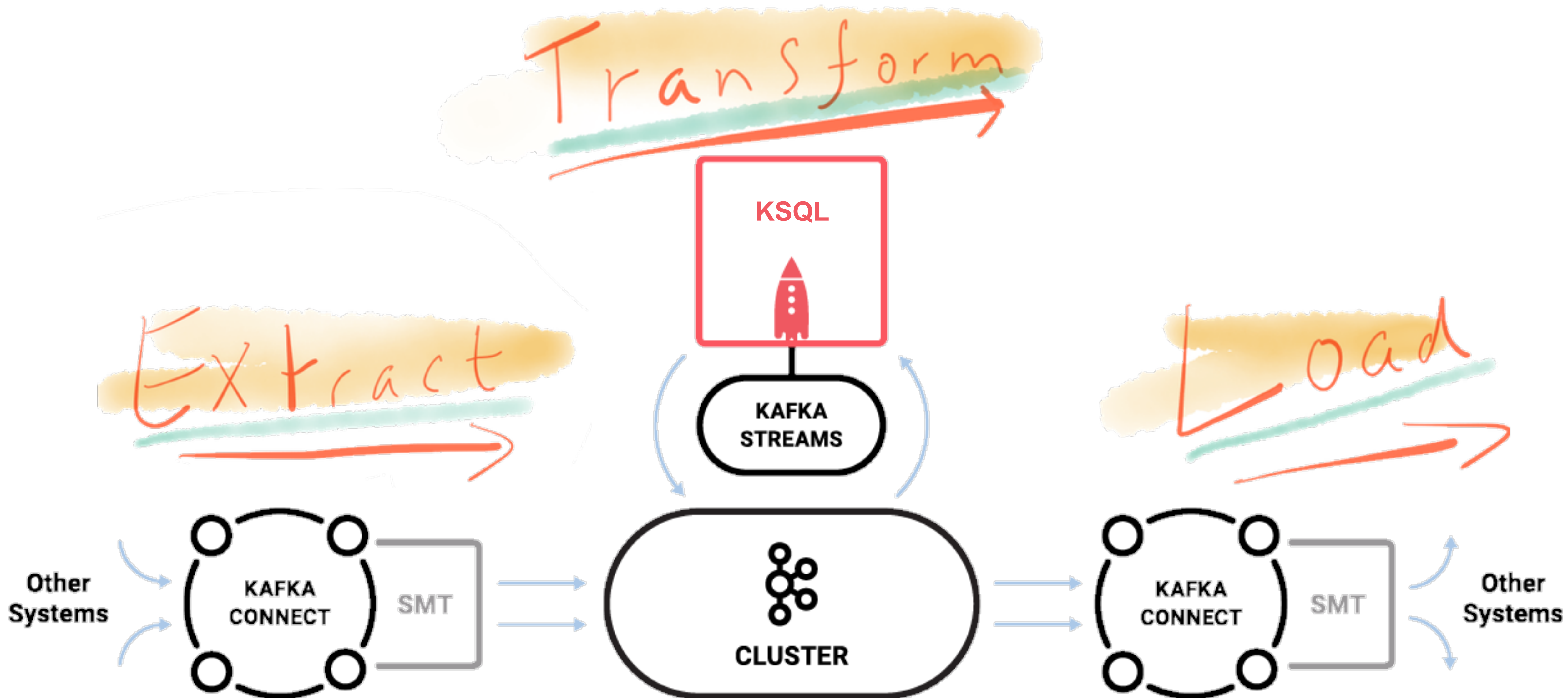
- Log data monitoring, tracking and alerting
- syslog data
- Sensor / IoT data

```
CREATE STREAM SYSLOG_INVALID_USERS AS  
SELECT HOST, MESSAGE  
FROM SYSLOG  
WHERE MESSAGE LIKE '%Invalid user%';
```

<http://cnfl.io/syslogs-filtering> / <http://cnfl.io/syslog-alerting>

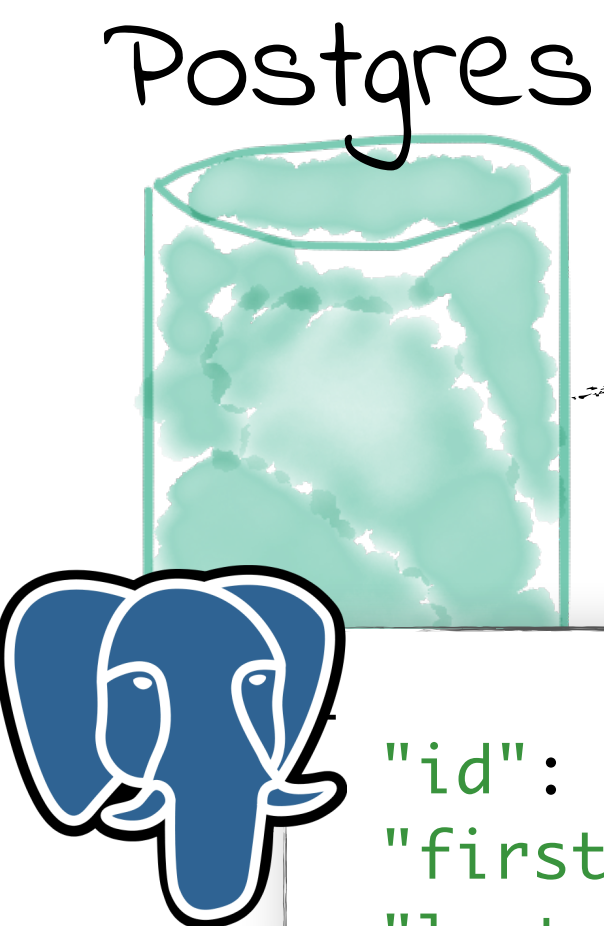


Streaming ETL with Apache Kafka



Demo Time!

```
{  
  "rating_id": 5313,  
  "user_id": 3,  
  "stars": 4,  
  "route_id": 6975,  
  "rating_time": 1519304105213,  
  "channel": "web",  
  "message": "worst. flight. ever. #neveragain"  
}
```

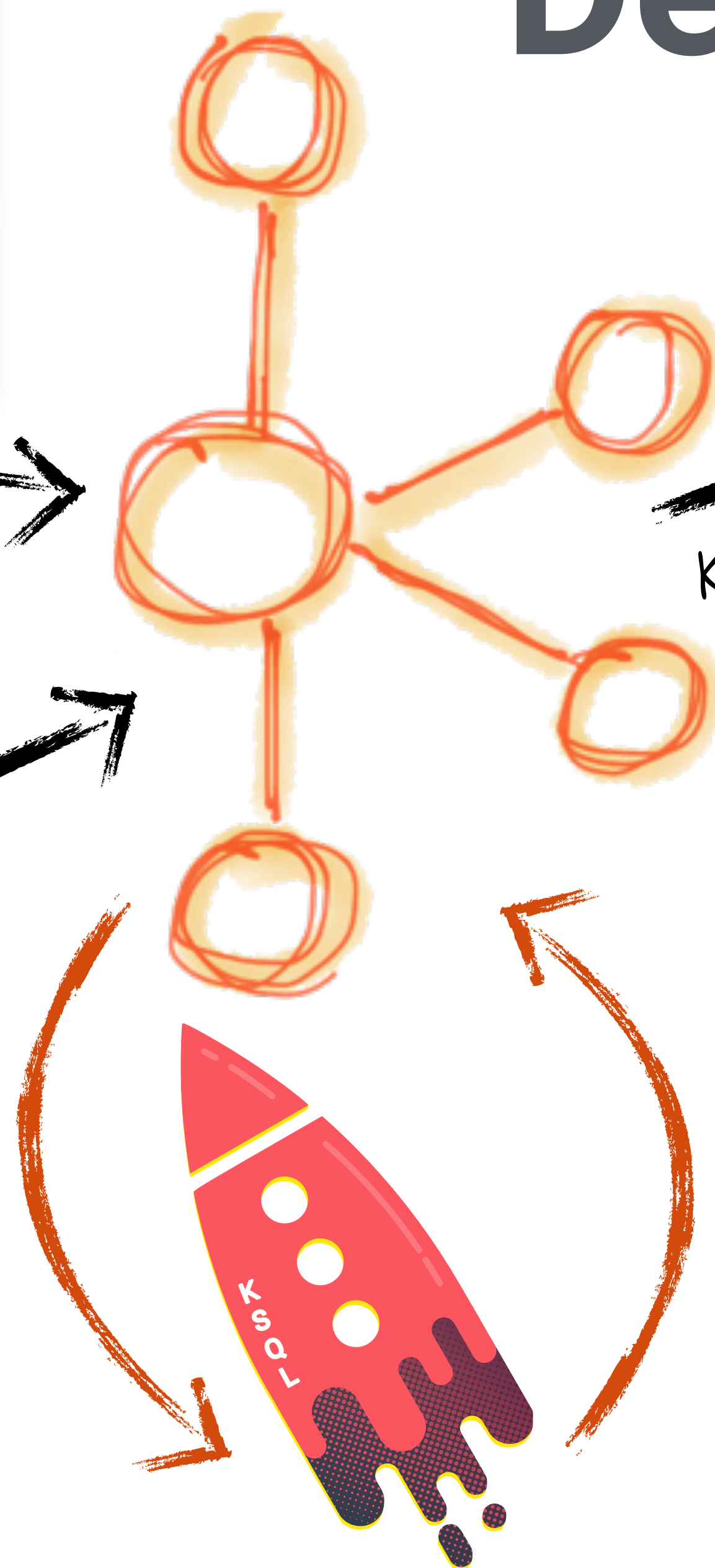


```
"id": 3,  
"first_name": "Marilyn",  
"last_name": "Doughartie",  
"email": "mdoughartie1@dedecms.com",  
"gender": "Female",  
"club_status": "platinum",  
"comments": "none"
```



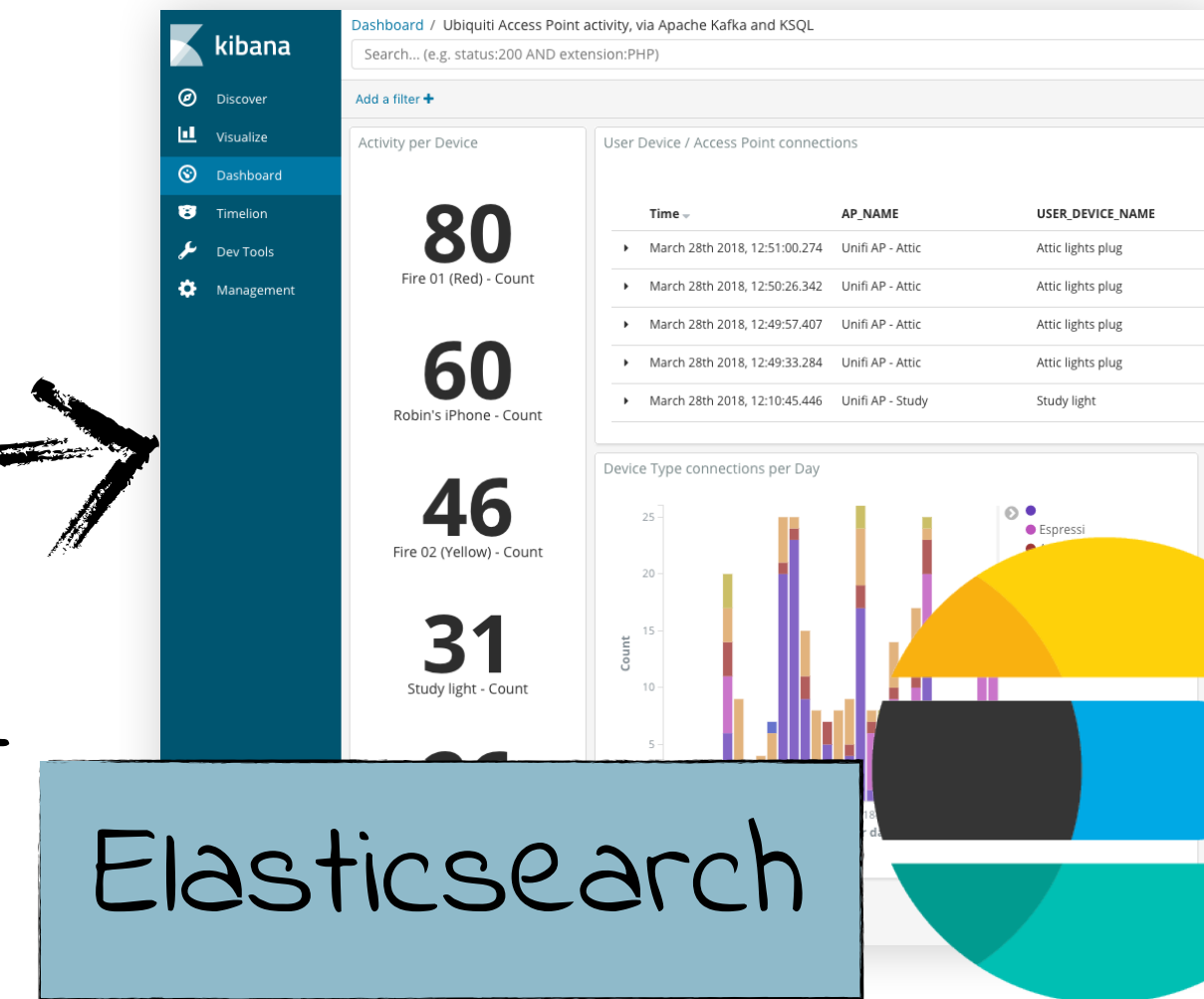
Kafka Connect

Producer API



Kafka Connect

Kafka Connect

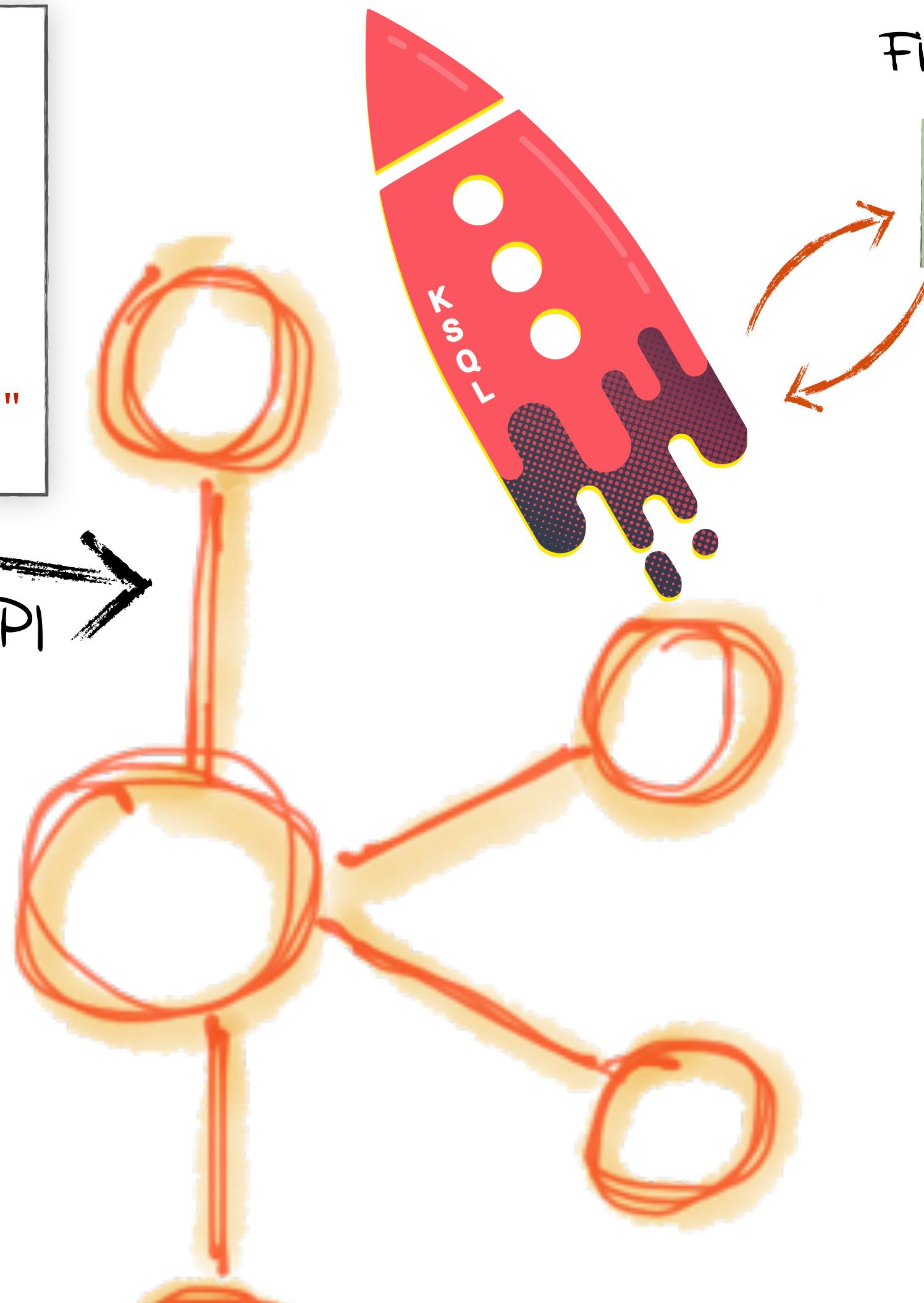


Elasticsearch

Postgres

```
{  
  "rating_id": 5313,  
  "user_id": 3,  
  "stars": 4,  
  "route_id": 6975,  
  "rating_time": 1519304105213,  
  "channel": "web",  
  "message": "worst. flight. ever. #neveragain"  
}
```

Producer API



Filter all ratings where STARS < 3

POOR_RATINGS

```
CREATE STREAM POOR_RATINGS AS  
SELECT * FROM ratings WHERE STARS < 3
```

```
{
  "rating_id": 5313,
  "user_id": 3,
  "stars": 4,
  "route_id": 6975,
  "rating_time": 1519304105213,
  "channel": "web",
  "message": "worst. flight. ever. #neveragain"
}
```

Producer API

Kafka Connect

Join each rating to customer data

RATINGS_WITH_CUSTOMER_DATA

Filter for just PLATINUM customers

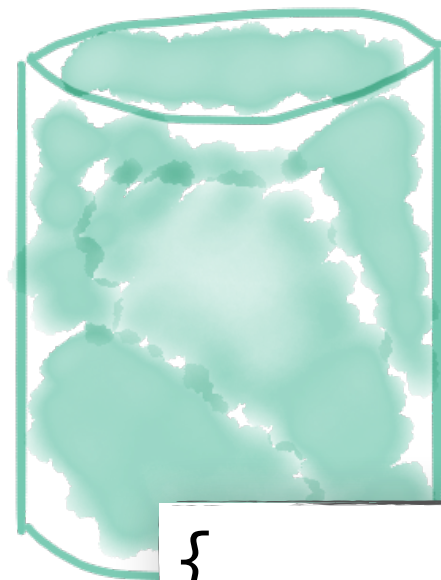
UNHAPPY_PLATINUM_CUSTOMERS

```
{
  "id": 3,
  "first_name": "M",
  "last_name": "M",
  "email": "m",
  "gender": "M",
  "club_status": "PLATINUM",
  "comments": "M"
}
```

```
CREATE STREAM UNHAPPY_PLATINUM_CUSTOMERS AS
SELECT * FROM RATINGS_WITH_CUSTOMER_DATA
WHERE STARS < 3
```

```
{
  "rating_id": 5313,
  "user_id": 3,
  "stars": 4,
  "route_id": 6975,
  "rating_time": 15193041052,
  "channel": "web",
  "message": "worst. flight."
}
```

```
CREATE TABLE RATINGS_BY_CLUB_STATUS AS
SELECT CLUB_STATUS, COUNT(*)
FROM RATINGS_WITH_CUSTOMER_DATA
WINDOW TUMBLING (SIZE 1 MINUTES)
GROUP BY CLUB_STATUS;
```



Kafka Connect

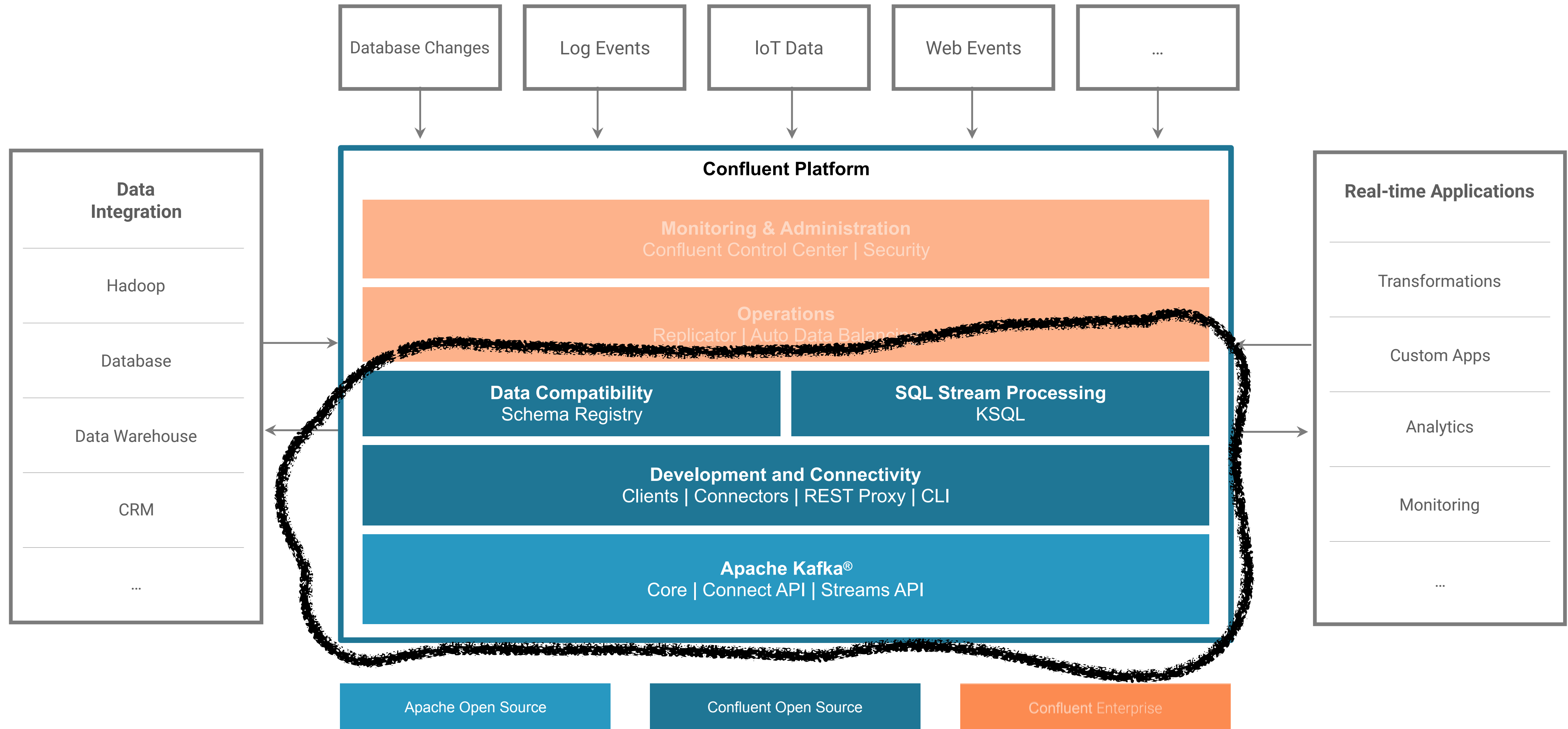
```
{
  "id": 3,
  "first_name": "Merilyn",
  "last_name": "Doughartie",
  "email": "mdoughartie1@dedecms.com",
  "gender": "Female",
  "club_status": "platinum",
  "comments": "none"
}
```



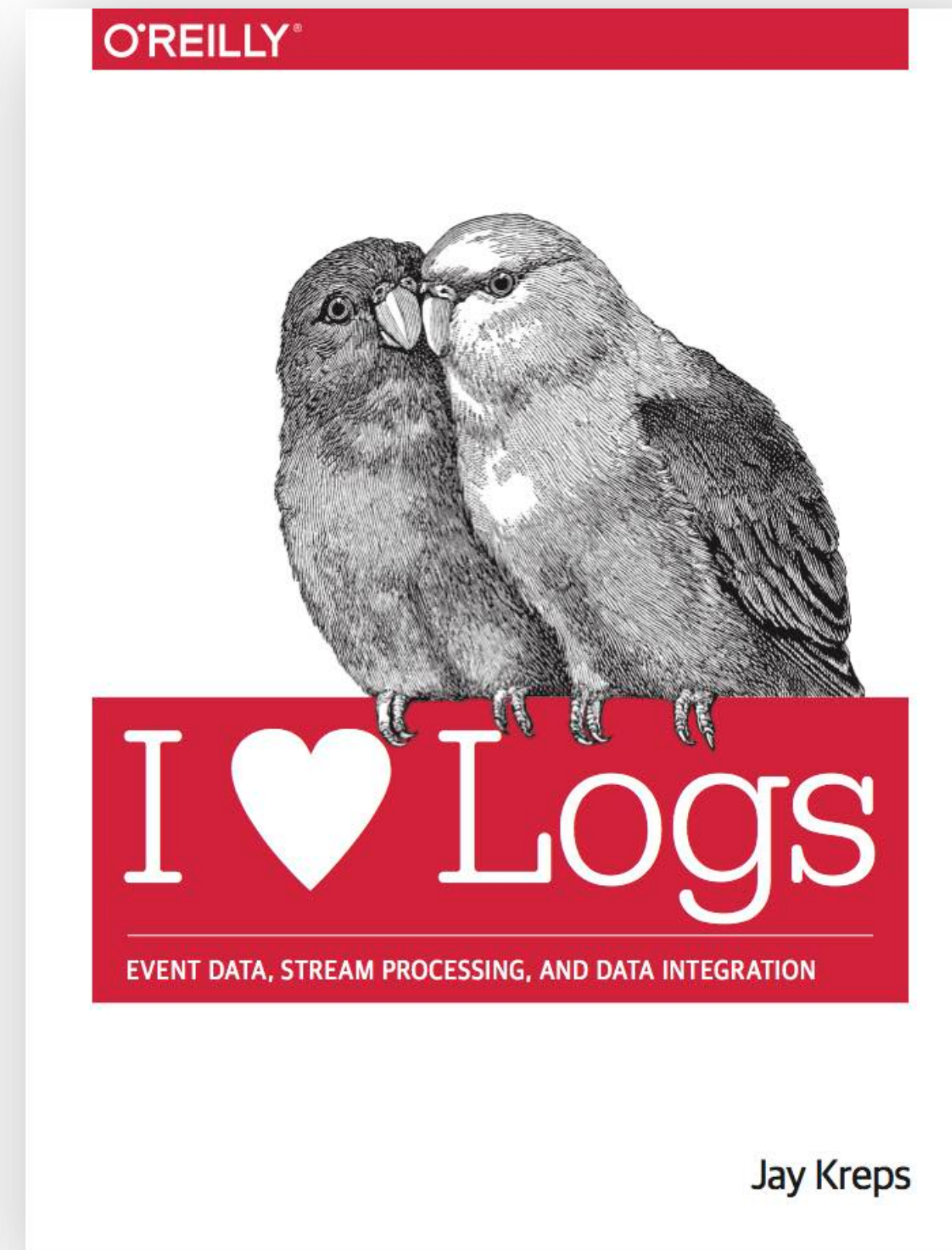
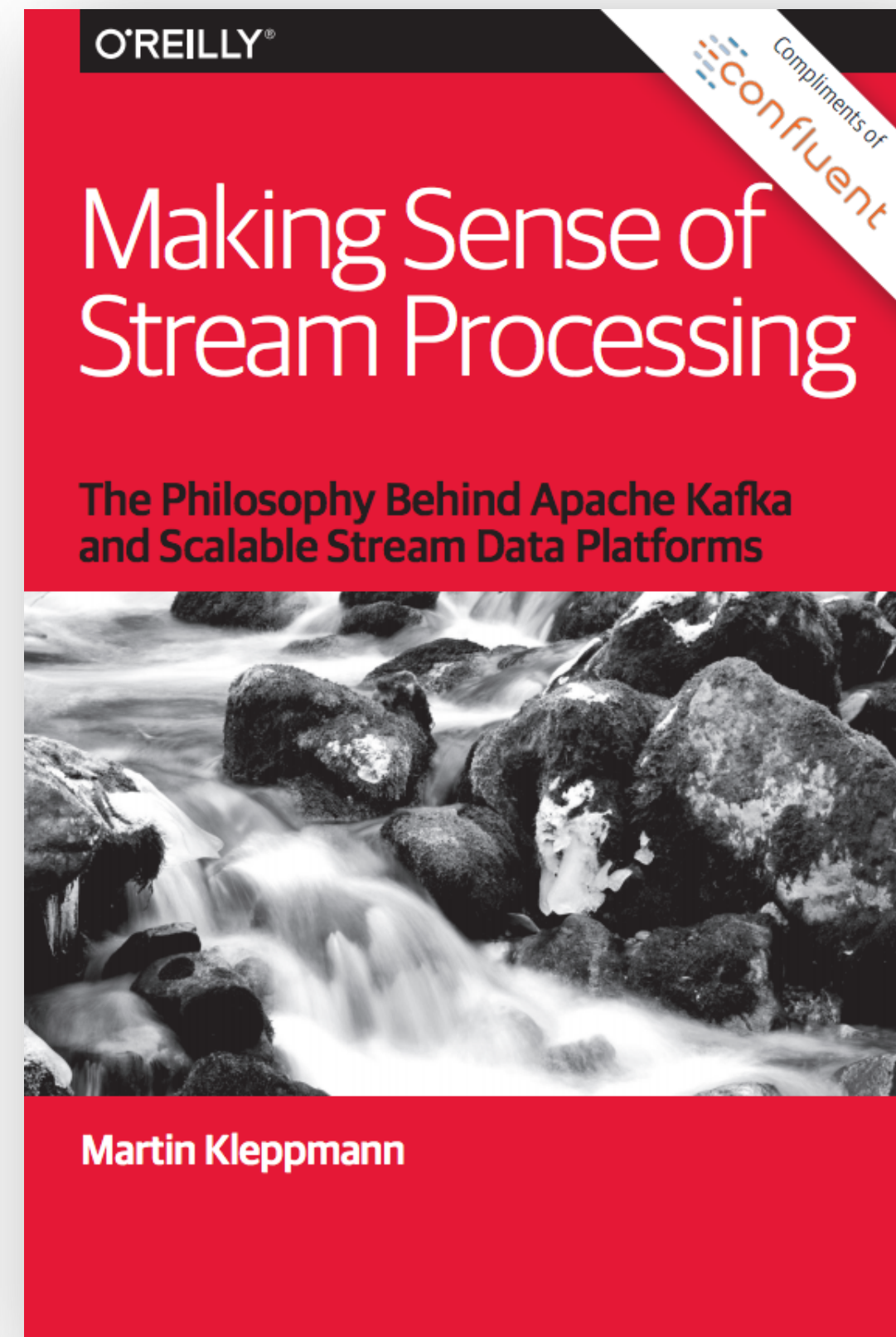
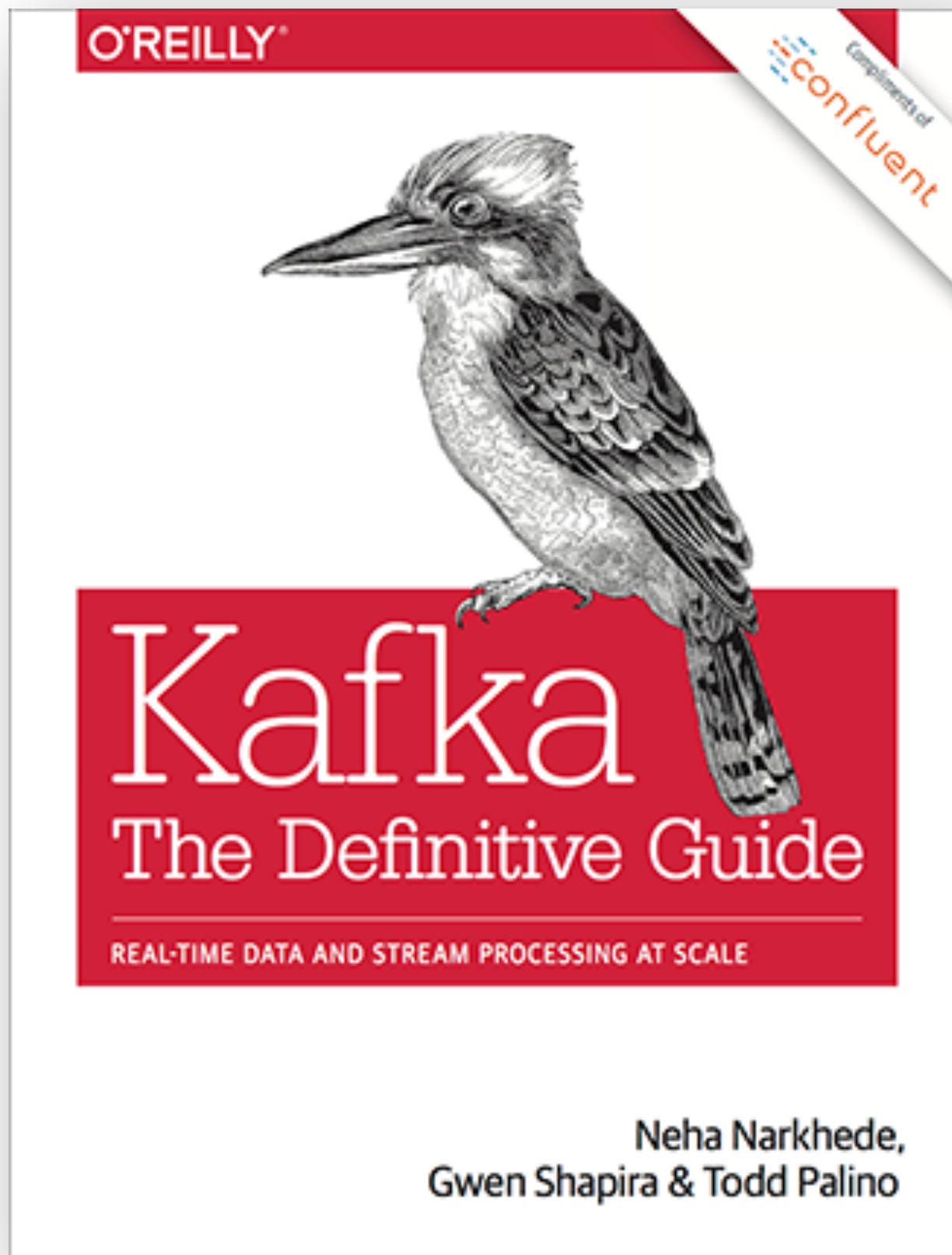
Aggregate per-minute by CLUB_STATUS

RATINGS_BY_CLUB_STATUS_1MIN

Confluent Open Source : Apache Kafka with a bunch of cool stuff! For free!



Free Books!



<https://www.confluent.io/apache-kafka-stream-processing-book-bundle>

<https://www.confluent.io/download/>

<http://cnfl.io/slack>



@rmoff robin@confluent.io

Useful links

- **Postgres integration into Kafka**

- <http://debezium.io/docs/connectors/postgresql/>
- <https://www.simple.com/engineering/a-change-data-capture-pipeline-from-postgresql-to-kafka>
- <https://www.slideshare.net/JeffKlukas/postgresql-kafka-the-delight-of-change-data-capture>
- <https://blog.insightdatascience.com/from-postgresql-to-redshift-with-kafka-connect-111c44954a6a>

- **Streaming ETL**

- Embrace the Anarchy : Apache Kafka's Role in Modern Data Architectures [Recording](#) & [Slides](#)
- [Look Ma, no Code! Building Streaming Data Pipelines with Apache Kafka and KSQL](#)
- Steps to Building a Streaming ETL Pipeline with Apache Kafka and KSQL [Recording](#) & [Slides](#)
- <https://www.confluent.io/blog/ksql-in-action-real-time-streaming-etl-from-oracle-transactional-data>
- <https://github.com/confluentinc/ksql/>

#EOF